UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Enhanced Depth Map in Low Light Conditions for RGB Cameras

Permalink

https://escholarship.org/uc/item/4nk4x934

Author

Chang, Joseph Daniel

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Enhanced Depth Map in Low Light Conditions for RGB Cameras

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Joseph D. Chang

Committee in charge:

Professor Truong Nguyen, Chair
Professor Cheolhong An
Professor Nikolay Atanasov

2021

The thesis of Joseph D. Chang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my Dad and Mom.

EPIGRAPH

They are all straight to him who understands,
and right to those who find knowledge.
—Proverbs 8:9 (ESV)

TABLE OF CONTENTS

Thesis Approval Page……..…..… iii

Dedication…… iv

Epigraph…….……. v

Table of Contents……..…… vi

List of Figures…….…….…… vii

List of Tables……... ix

Vita…….…….…… x

Abstract of the Thesis...…….…….……. xi

Chapter 1 Introduction.…….…… 1

Chapter 2 Background and Overview…….……. 4

Chapter 3 Datasets…….……... 15

Chapter 4 Enhanced Depth Map for Low-Light…...…… 18

Chapter 5 Enhanced Object Detection for Low-Light……. 36

Chapter 6 Conclusion…….……..……. 45

Bibliography …….…….…… 46

LIST OF FIGURES

Figure 1: Passive Stereo Matching with two cameras to calculate depth. ……..…… 5

Figure 2: Active Stereo Vision with two cameras and projector to calculate depth. …. 6

Figure 3: Design of CETM method. ……..….. 9

Figure 4: Nighttime images before and after CETM. ……..… 10

Figure 5: Single frame from turbulent video with heat haze before and after ……. 11
CETM.

Figure 6: Pretraining daytime image model f_D. ……..... 12

Figure 7: GAN training cycle for daytime and nighttime images. ……..... 13

Figure 8: Input images and output disparity predictions. …….……. 14

Figure 9: Oxford dataset color conversion and sparse ground truth conversion. ..…. 15

Figure 10: Sample data 01, 02, and 03 selected from the custom dataset with 714 …… 17
images. Shown are the (a) normal nighttime images, (b) enhanced images run through CETM, and (c) ground truth bounding box coordinates of objects detected.

Figure 11: Inverted YUV nighttime image looks like foggy daytime image. …..…...… 20

Figure 12: Preprocessing algorithm utilizing CETM defogging. ……..…….. 20

Figure 13: Nighttime images before and after preprocessing. …… 20

Figure 14: Estimated disparity for normal nighttime images and preprocessed …... 22
nighttime images for stereo camera using Stereo-Consistent Cyclic Translations [10].

Figure 15: (a) Accumulated pixel percentage accuracy for stereo depth ..…. 25
estimation with pretrained model on 6,001 nighttime images in the Oxford dataset without cropping the front car hood. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.

Figure 16: (a) Accumulated pixel percentage accuracy for stereo depth ...… 26
estimation with pretrained model on 6,001 nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.

Figure 17: Estimated disparity for normal nighttime images, preprocessed ...… 28

nighttime images, and sparse ground truth for monocular camera using monodepth2 [3].

Figure 18: Estimated disparity for normal nighttime images, preprocessed nighttime images, and sparse ground truth for stereo cameras using Stereo-Consistent Cyclic Translations [10]. ……. 29

Figure 19: (a) Accumulated pixel percentage accuracy for monocular depth estimation with pretrained model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA. …… 31

Figure 20: (a) Accumulated pixel percentage accuracy for stereo depth estimation with pretrained model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA. …… 32

Figure 21: Test custom disparity estimation model with nighttime images. ……..…… 34

Figure 22: Estimated disparity on custom model with nighttime, preprocessed nighttime images, and sparse ground truth using Stereo-Consistent Cyclic Translations [10]. ..…. 34

Figure 23: (a) Accumulated pixel percentage accuracy for stereo depth estimation with custom model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA. ...… 35

Figure 24: Objects detected in normal nighttime images and preprocessed nighttime images using YOLOv5 pretrained model. …... 37

Figure 25: Mean average precision (mAP) evaluation on YOLOv5 nighttime object detection for pretrained model with (a) IOU=0.5 (b) IOU=0.7 (c) IOU=0.85 (d) IOU=0.93 …… 40

Figure 26: Objects detected in normal nighttime images and preprocessed nighttime images using YOLOv5 self-trained model …... 42

Figure 27: Mean average precision (mAP) evaluation on YOLOv5 nighttime object detection for self-trained model with (a) IOU=0.5 (b) IOU=0.7 …… 44

LIST OF TABLES

Table 4.1: Summary of User Driving Test Results …… 19

VITA

2015-2020	B. S. in Electrical Engineering, University of California San Diego
2020-2021	M. S. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego

ABSTRACT OF THE THESIS

Enhanced Depth Map in Low Light Conditions for RGB Cameras

by

Joseph D. Chang

Master of Science in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2021

Professor Truong Nguyen, Chair

Most existing depth estimation methods predict depth for daytime images and do not perform well in low-light situations due to lack of clear environmental features, glare, overexposure, and noise. This is problematic for safe autonomous driving as pedestrian and guardrail detection at night is challenging and poses potentially life-threatening situations. This thesis addresses this problem by improving image quality of disparity maps obtained in low-light

using a preprocessing method. We introduce an algorithm that combines a defogging method, which enhances night images and improves luminance, with generative adversarial or fully-convolutional networks to accurately learn the correct disparity prediction. The experiments show that the proposed method outperforms state-of-the-art methods which do not use additional preprocessing.

Chapter 1

Introduction

Depth estimation based on stereo vision is a major topic in computer vision and is key to enhancing safety in advanced driver assistance systems (ADAS). It is used to gain 3D understanding of the world from 2D images [1][2]. Although monocular depth estimation has lower accuracy than stereo-based depth estimation, it would enable many 3D datasets in the future since the depth of any image can be computed [3][4]. Another popular method is to use Lidar to estimate depth by measuring a laser's return time from its target. Although Lidar-based methods are highly accurate, they require expensive equipment for longer-range applications such as driving scenarios [5]. Hence, stereo RGB cameras are currently the best method for depth estimation as they produce accurate depth while not requiring expensive sensors. Despite the importance of depth estimation to existing applications such as virtual reality, autonomous cars, and robotic surgery, the low-light and nighttime scenarios have hardly been explored which restricts these systems to run only in daytime or under artificial light.

Low-Light depth estimation is an important problem as vehicles can never be fully autonomous unless they can operate in twilight or at night. Safety at night is essential as most accidents occur between midnight and 4:00 a.m. According to a traffic safety study in May 2007, the passenger vehicle occupant fatality rate at nighttime is about three times higher than the daytime rate. Unrestrained passengers experience 17% higher fatality rate at night than daytime.

At night, 3.3 times more accidents were attributed to alcohol involvement than daytime. In addition, 16% more fatalities occur from speeding at night [6]. These statistics show the proportion of nighttime crash fatalities among passenger vehicle occupants is higher compared to daytime. Additionally, people suffer from various health conditions that impede driving at night. One example is night blindness (nyctalopia) which causes blurry vision while driving in the dark due to inability to quickly adapt from a well-illuminated to poorly-illuminated environments [7]. For complete ADAS, object detection and segmentation algorithms must be improved through higher accuracy depth estimation to account for low-light.

In this paper, we formulate an algorithm to improve estimated depth map quality in low-light conditions for ADAS by adding a preprocessing method to existing unsupervised learning algorithms. Since there are nearly no 3D nighttime street view datasets with ground truth available, supervised learning is unsuitable. Oxford RobotCar provides nighttime car datasets with sparse depth from Lidar. Hence, our method utilizes these datasets to test if the proposed method accurately generates a disparity map given any nighttime RGB image. Then, the data is tested with monocular and stereo depth estimation algorithms with and without our preprocessing enhancement. The evaluation metrics show that the proposed method outperforms the current state-of-the-art methods.

Additionally, we train a custom model using these preprocessed nighttime images for the network to select more features and generate disparity maps with greater accuracy given any nighttime RGB images. The model is tested with various nighttime images and evaluation metrics which show that the proposed method outperforms current state-of-the-art methods in low-light depth estimation. In future works, the enhanced depth maps from our algorithm can be applied to object detection and classification for autonomous vehicles driving at night. To spearhead the

object detection application, we created a dataset from Berkeley Deep Drive which contains nighttime monocular images, corresponding preprocessed images, object detection ground truth labels, and a list of classes which can be tested with state-of-the-art object detection algorithms such as You-only-look-once (YOLO) [8] for real time object detection. We utilize YOLOv5 to detect objects in nighttime street view images and evaluate with the mAP (mean average precision) metric to show that the proposed method improves object detection.

Chapter 2

Background and Overview

2.1 Background in Depth Estimation

Depth estimation has been studied since the 1990s and was based on the way the human vision system determines depth of objects from 2D images. The brain uses depth cue signals to calculate how far away an object is. Some depth cue examples are that closer objects are typically larger, closer objects have high-quality texture while further objects are blurry, and lines such as road lanes tend to converge at the horizon helping determine the depth of objects along the lines. Because the human vision system has two eyes, it experiences stereopsis where closer objects have greater disparity, specifically they move further horizontally when viewed out of each individual eye. The brain translates this larger disparity to closer objects and smaller disparity to further objects. This is the biological basis for stereo-camera depth estimation.

Before deep learning, scientists used passive stereo vision (PSV) [31] estimates depth by exploiting epipolar geometry. The first step is camera calibration where feature points in both images are identified and matched. The camera intrinsics, extrinsics, and lens distortion parameters are estimated. The second step is rectification where the stereo images are projected onto a common epipolar plane, so they are not warped. Finally, the disparity is calculated as $d = x_1 - x_2$ shown in Figure 1. Depth is derived using similar triangles as $z = fb/d$ where f is focal length, b is baseline distance between the cameras, and d is disparity. This method performs well in sunlight and is cost-effective. However, it has mediocre performance in low-light and non-textured scenes.

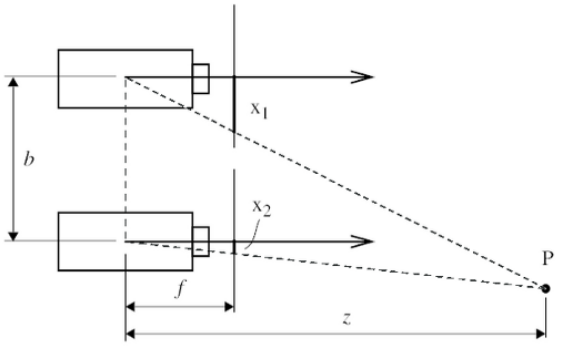


Figure 1: Passive Stereo Matching with two cameras to calculate depth.

Over time, more advanced forms of stereo vision were developed to include laser or structured light to simplify the stereo matching problem as shown in Figure 2. Conventional structured-light vision (SLV) [32] uses a structured light or laser to project a known pattern of grids or bars onto a target surface. The way patterns deform allows the system to find the projector-camera correspondences and calculate depth. Conventional active stereo vision (ASV) also uses structured light or laser to project texture onto surfaces. This mitigates the need for an external light source to find camera-camera correspondences and calculates depth. It is useful in low-light and non-textured indoor scenes. However, it performs poorly under sunlight and over longer distances. The added infrared projector also increases cost. Structured-light stereo (SLS) is a hybrid method that uses both projector-camera and camera-camera correspondences to calculate the most accurate depth.

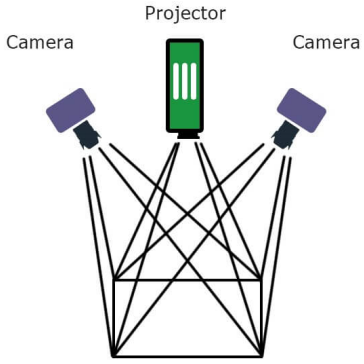


Figure 2: Active Stereo Vision with two cameras and projector to calculate depth.

2.1.1 Monocular Depth Estimation

As hardware became faster and new machine learning techniques were developed, deep learning took over much depth estimation research. Because stereo depth estimation has been heavily studied, most recent work focuses on monocular depth estimation with datasets such as KITTI [22] and NYUv2 [23]. This began with Eigen et al. [24] in 2014 who used a CNN to train the first monocular depth estimation model with training images and ground truth. It used two deep network stacks: one to make a coarse global prediction on the entire image and another to refine the prediction in local sections. It achieved state-of-the-art results on NYU Depth and KITTI without superpixelation.

In 2016, Garg et al. [4] proposed the first monocular depth estimation algorithm to learn depth without ground truth. To train, it considers a pair of images called source and target with small, known camera motion between the pair. A convolutional encoder is trained to predict depth of the source image. This requires generating an inverse warp of the target image using predicted depth and displacement between the two viewpoints to reconstruct the source image. The

photometric error in the reconstruction is loss for the encoder. The network trained on less than half of KITTI data to obtain comparable performance to state-of-the-art methods.

Godard et al. [3] further improved monocular depth estimation in 2016 and the method does not require ground truth depth either. Utilizing epipolar geometry, it generates a disparity image by training the network with image reconstruction loss. Alone, this produces low quality depth which is fixed by updating the loss to enforce consistency between the disparities produced by both the left and right images. It produced state-of-the-art results on the KITTI dataset and outperformed supervised methods trained with ground truth at the time.

In 2017, Zhou et al. [25] and Wang et al. [26] proposed methods unsupervised monocular depth estimation without ground truth. Zhou simultaneously trained depth and camera pose estimation networks using view synthesis as supervision. The networks were coupled by the view synthesis objective during training. Evaluation on KITTI showed that the method performed comparably with methods that use ground truth pose or depth for training. Wang argued that the depth CNN predictor can be learned without a pose CNN predictor at all. This is achieved by incorporating a differentiable implementation of direct visual odometry (DVO) and a novel normalization strategy. The results improved performance over state-of-the-art single-view depth.

2.1.2 Stereo Depth Estimation

Despite the progress in monocular depth estimation, there is still a large gap between it and stereo depth estimation. Stereo camera setups still generally yield lower error. In 2018, Smolyanskiy [27] proposed a semi-supervised deep learning algorithm to train a stereo model. The architecture includes a machine-learned argmax layer and custom runtime that enables a smaller

version of the stereo network to run on an embedded GPU. It was tested on the KITTI 2015 stereo dataset and produced competitive results.

In 2018, Khamis [28] presented StereoNet, an end-to-end deep architecture for real-time stereo matching at 60 fps on a Nvidia Titan X. It achieves sub-pixel matching precision allowing real-time performance with a very low resolution cost volume that encodes all information needed for high disparity precision. Spatial precision is achieved by employing a learned edge-aware upsampling function. The model uses a Siamese network to extract features from the left and right images. The first disparity estimate computed is a low resolution cost volume. The model then reintroduces high frequency details through a learned upsampling function that uses compact pixel-to-pixel refinement networks. It achieves compelling results on Scene Flow, KITTI 2012, and KITTI 2015 benchmarks and is extremely flexible for an acceptable computational cost.

Tankovich [29] proposed HITNet, a neural network for real-time stereo matching in 2021. Most recent multi-view stereo approaches are based on a 3D cost-volumes and 3D convolutions which are memory consuming and not feasible for high resolution reconstruction. However, this method is able to run quickly by relying on fast multi-resolution initialization and differentiable 2D geometric propagation of information across different levels. In addition, warping mechanisms infer disparity. It ranked 1st-3rd on all two-view stereo metrics published on the ETH3D website and ranks 1st on most metrics among all end-to-end learning approaches on Middleburv3. It ranks 1st in KITTI 2012 and 2015 benchmarks among published methods faster than 100ms per image.

In 2021, Huang [30] developed the Efficient Stereo Network (ESNet). Typically, 3D convolutions are used to improve performance of existing networks, but they are slow, computationally expensive, and would not work in real-time applications. However, ESNet still relies only on 2D convolutions and computes multi-scale cost volume efficiently using a warping

method to improve performance in regions with fine-details. It further improves performance by addressing matching ambiguity in occluded regions with an occlusion mask estimation network ESNet-M and a new training scheme that includes dataset scheduling and unsupervised pre-training. The method achieved state-of-the-art on Scene Flow, DrivingStereo, and the KITTI 2015 stereo dataset.

2.2 Contrast Enhancement Turbulence Mitigation

In recent decades, there has been much development in image enhancement to increase visibility in foggy areas such as near oceans. The major problems to tackle include fog, haze, and turbulence. When fog is present, images captured by cameras have low contrast. It is difficult to identify edges between objects and their background. Turbulence from high temperature adds heat haze which blurs an image resulting in a less clear image. The Contrast Enhancement Turbulence Mitigation (CETM) [9] method jointly enhances contrast while also performing turbulence removal. Here, we provide an overview of the method which is shown in Figure 3.

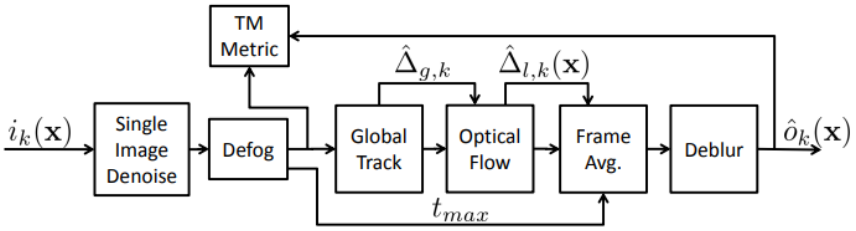


Figure 3. Design of CETM method.

The input to the system $i_k(x)$ is a foggy, turbulent image at frame k. It is passed through a single image denoising and defogging algorithm to smooth out image noise. The defogging step estimates the transmission and t_{max} and enhances the contrast image. Then, an object tracker

estimates the global motion vector $\hat{\Delta}_{\mathrm{g,k}}$. That enhanced image is aligned with $\hat{\Delta}_{\mathrm{g,k}}$ and locally aligned with an optical flow method which estimates $\hat{\Delta}_{l,k}(x)$. Next, the globally and locally aligned contrast enhanced image is temporally averaged with *K* frames. Finally, the deblur step removes atmospheric distortion resulting in an enhanced image $\hat{o}_k(x)$. Low-light images before and after CETM enhancement in Figure 4 show significantly brighter images with clearer features. A turbulent video of cars driving through a hot desert contains heat haze that is mitigated after CETM enhancement as shown in Figure 5. The performance of the CETM is measured using the turbulence mitigation metric (TMM) by comparing the turbulence between the input images after denoise and defog with the output images.

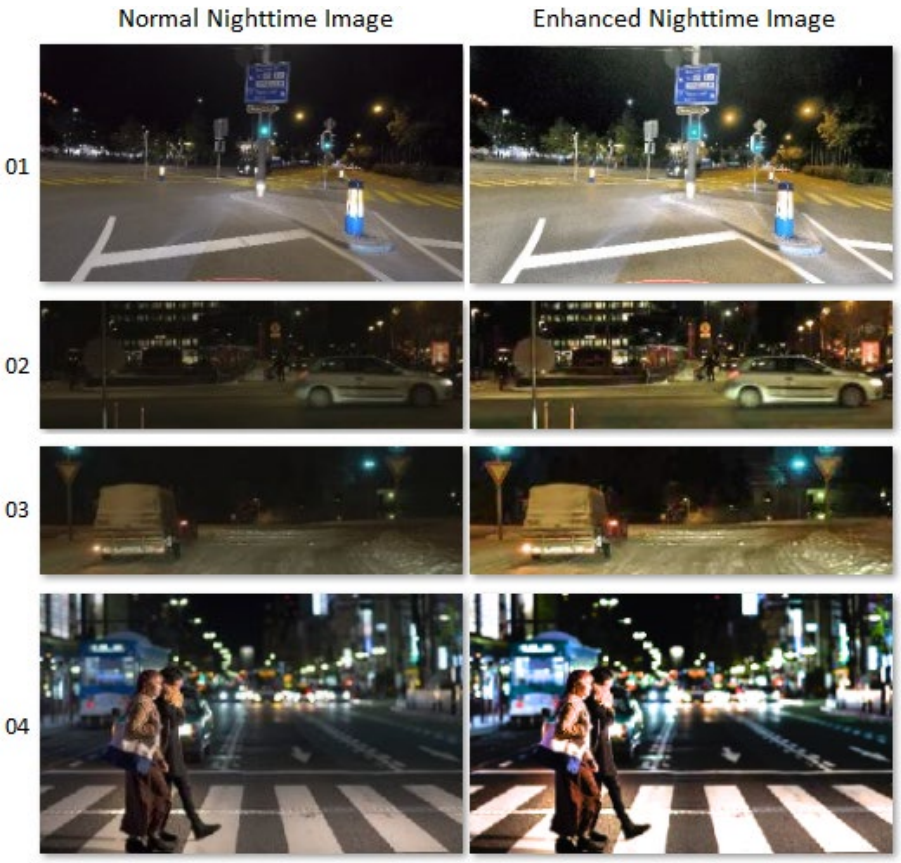


Figure 4. Nighttime images before and after CETM.

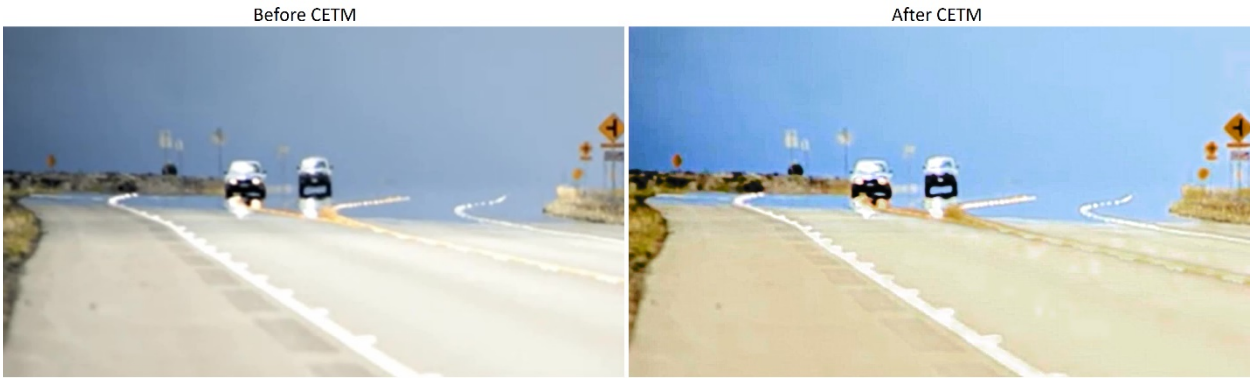


Figure 5. Single frame from turbulent video with heat haze before and after CETM.

2.3 Low-Light Depth Estimation Algorithm

A recent study by Sharma et al. [10] introduced a joint translation and stereo network to solve low-light depth estimation. The solution is an unsupervised generative adversarial network (GAN) [11] that does not require ground-truth disparities to train. The idea is to train a translation network that transforms stereo nighttime images to rendered stereo daytime images and another model that transforms the original stereo daytime images to rendered stereo nighttime images. Simultaneously, a stereo network is trained to compute the disparity map of the rendered nighttime image pairs. This is guided by the disparity map computed from the original daytime image pair. Here, we provide additional details of the algorithm in the subsection below.

2.3.1 Training

It begins by pretraining weights for daytime depth estimation using stereo daytime images X_{LD}, X_{RD} and their corresponding ground truth Y_{LD}, as shown in Figure 6. The images are passed through a stacked hourglass CNN model for training which gives the daytime disparity estimate weights f_D.



Figure 6. Pretraining daytime image model f_D.

Next, the images are run through two separate training cycles. Each training cycle consists of a translation network [14] and a stereo network modeled after PSMNet [15]. In the first cycle, stereo daytime images X_{LD}, X_{RD} shown in Figure 7 are passed through a daytime generator g_D that renders an estimate of corresponding stereo nighttime images $\hat{X}_{LN}, \hat{\mathrm{X}}_{RN}$. $\hat{\mathrm{X}}_{LD}, \hat{\mathrm{X}}_{RD}$ and $\hat{X}_{LN}, \hat{\mathrm{X}}_{RN}$ are passed through the pretrained daytime weights f_D and nighttime weights f_N initialized as f_D, respectively. Stereo-consistency then computes the similarity of the two disparity estimates and updates f_N and g_D with its discriminator d_D. In addition, $\hat{X}_{LN}, \hat{\mathrm{X}}_{RN}$ are passed through a nighttime generator g_N which renders stereo daytime images X_{LD}, X_{RD} for the second training cycle.

In the second training cycle, stereo nighttime images $\mathrm{X}_{LN}, \mathrm{X}_{RN}$ shown in Figure 6 are passed through a nighttime generator g_N which renders corresponding daytime images $\hat{\mathrm{X}}_{LD}, \hat{\mathrm{X}}_{RD}$ is passed through a daytime generator to produce rendered nighttime images $\tilde{\mathrm{X}}_{LN}, \tilde{\mathrm{X}}_{RN}$. $\hat{\mathrm{X}}_{LD}, \hat{\mathrm{X}}_{RD}$ and $\tilde{\mathrm{X}}_{LN}, \tilde{\mathrm{X}}_{RN}$ are passed through the f_D from pretraining and f_N from the first cycle, respectively. Then, stereo-consistency compares the similarity of the disparity estimates and updates the weights of g_N and its discriminator d_N. This process iterates for as many epochs as specified until weights

$\hat{g}_N^*$, $\hat{g}_D^*$, and $\hat{f}_N^*$ are optimized. To test nighttime images X_{LN}, X_{RN} we pass them through $\hat{g}_N^*$, $\hat{g}_D^*$, and $\hat{f}_N^*$ resulting in estimated disparity $\hat{y}_{LN}^*$.

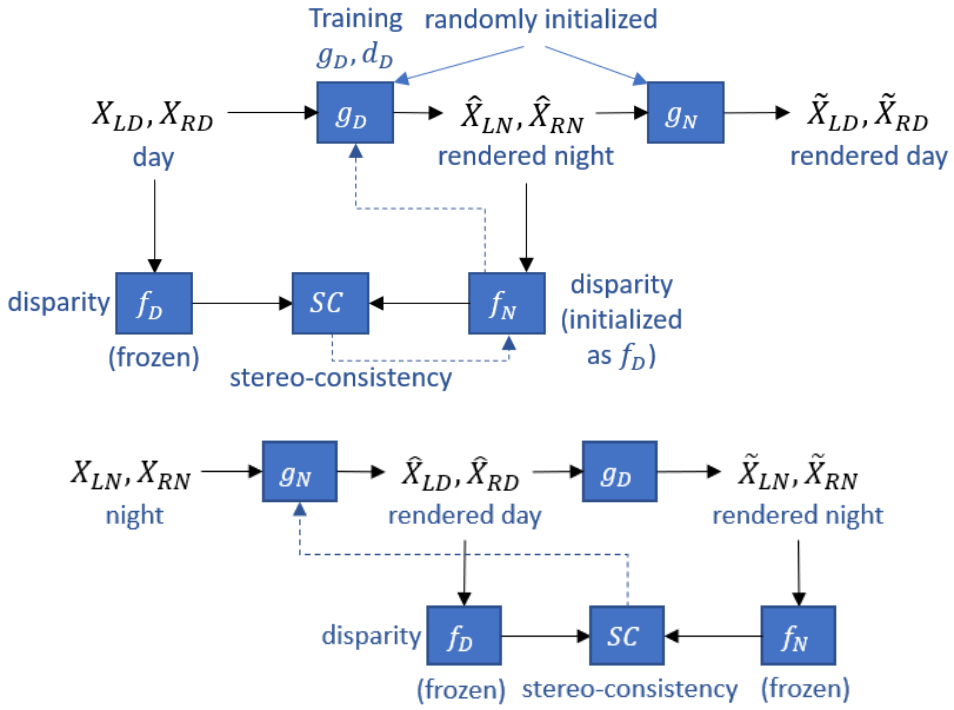


Figure 7. GAN training cycle for daytime and nighttime images.

2.3.2 Network Architecture

Each translation network is a GAN containing a daytime and nighttime generator g_D and g_N to predict disparity. Each generator contains an encoder with three convolution layers and four residual blocks [16] followed by a decoder with four residual blocks, two deconvolution layers, and one convolution layer. Each translation network also has daytime and nighttime discriminators d_D and d_N to classify prediction accuracy. Each discriminator contains three sub-discriminators that are 32×32 Patch GANs [17] with five convolution layers. The respective inputs to each sub-

discriminator are a RGB image blurred with a 5×5 Gaussian kernel, a grayscale image containing only the luminance channel, and an image containing horizontal and vertical gradients [14].

The stereo networks for daytime and nighttime weights f_D, f_N based on the PSMNet [15] CNN contain three convolution layers, four residual blocks, and a Spatial Pyramid Pooling [18] module. The extracted features are concatenated into a 4D cost volume then regularized by a 3D convolution architecture. Lastly, the features are passed through a disparity regression [19] layer to obtain the final disparity prediction shown in Figure 8.



Figure 8. Input images and output disparity predictions.

Chapter 3

Datasets

3.1 Oxford Dataset

We use the Oxford RobotCar Dataset [12][13] which contains a mix of daytime and nighttime street view data in various weather conditions. The images are stored in a raw format, so we convert them to RGB using a lookup table (LUT). The data is then ready for preprocessing and training. However, we also need sparse ground truth disparity to calculate error later. For this, we gather the camera model and extrinsic parameters of the Bumblebee XB3 stereo camera, 2D Lidar data from the front bumper, Inertial Navigation System (INS) data, and timestamps from the Lidar and stereo camera for each dataset. The sparse Lidar depth maps are projected into the camera resulting in sparse depth maps which are converted to sparse disparity maps. As a result, we obtain a stereo dataset with ground truth disparity. The complete process is shown in Figure 9.

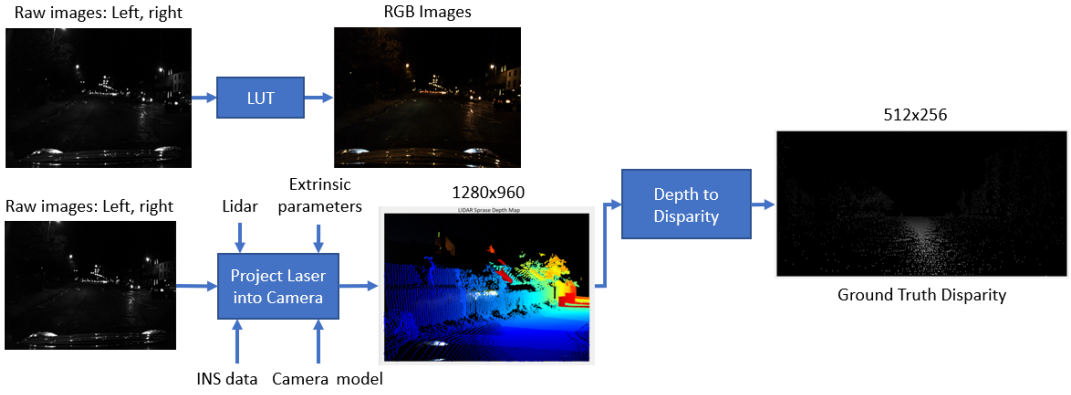


Figure 9. Oxford dataset color conversion and sparse ground truth conversion.

3.2 Berkeley Deep Drive Dataset

We also plan to perform object detection, but the Oxford dataset does not have ground truth labels for object detection. Hence, we use the Berkeley Deep Drive (BDD), the largest driving video dataset which contains 100k monocular street view images with object detection labels. Each image in the dataset is taken in different lighting and weather conditions making it very realistic for real-world applications. However, the dataset contains both daytime and nighttime images with more of the former. We manually filter out nighttime images to create our own custom datasets of 8k, 7130, 4159, and 714 images each with corresponding ground truth object detection files. The larger datasets contain a greater variety of nighttime images including those with motion blur, weird camera angles, dirty windshields, or very well-lit while the smaller datasets contain only darker, clearer street view images. Each dataset contains nighttime images which each have a corresponding image enhanced with CETM, a file defining which class numbers correspond to which objects, and individual files with object class numbers and bounding box coordinates as shown in Figure 10. We plan to publish our datasets as open source for future studies in nighttime object detection.



Figure 10. Sample data 01, 02, and 03 selected from the custom dataset with 714 images. Shown are the (a) normal nighttime images, (b) enhanced images run through CETM, and (c) ground truth bounding box coordinates of objects detected.

Chapter 4

Enhanced Depth Map for Low-Light

Given a RGB stereo image pair or monocular image of a nighttime scene, our goal is to calculate the disparity map where each pixel represents the distance from the stereo camera to a point in space. The proposed approach consists of two steps. First, it uses a defogging algorithm to preprocess and enhance the image. Then, it uses a neural network based approach such as CNN or GAN to estimate the disparity map.

4.1 Image Preprocessing Method

Given a YUV nighttime image I, we invert its luma channel with 255-I as shown in Figure 11 and notice the result looks similar to a foggy daytime image. We utilize this idea to construct a preprocessing procedure that enhances nighttime images by running them through a defogging process. We hypothesize that this process will improve image clarity and depth estimation. The improved contrast should make the images appear as if they were taken during twilight where objects have clearer edges and texture. This will give the networks more features to extract while training and lead to better disparity prediction on nighttime images. To defog our dataset, we choose the joint contrast enhancement and turbulence mitigation method (CETM) [9] shown in Figure 12 which is designed to mitigate turbulence and remove fog in images while being fast enough for real-time applications. First, we convert the nighttime images to the YUV domain and

then invert the luminance channel. Then, CETM is applied to denoise and defog the images with atmospheric blur reduction. The result is smoothed by feature tracking using optical flow. To restore the image, we un-invert the luma channel with 255-I and convert the image back to RGB. Samples of the results of nighttime images before and after preprocessing are shown in Figure 13. It is clear that after enhancement, nighttime images have greater visibility and object edges are clearer. To increase depth estimation accuracy, we enhance images before feeding them through depth estimation models. Since preprocessing method is performed on single images at a time, it can be used for both monocular and stereo applications. We think that the proposed method will yield more accurate depth estimation for both, which is useful since both will be widely used in the future.

Before further experiments, we create a subjective user driving test to check our hypothesis that CETM improves nighttime vision is in the right direction. The test has 14 users choose whether normal or enhanced images are clearer if driving for 100 nighttime images from the custom Berkeley Deep Drive dataset of 4159 images. The results are shown in Table 4.1 where enhanced images are 100% preferred for 59% of tests and 57.1% preferred for only 1% of tests.

Table 4.1: Summary of User Driving Test Results

Number of Images	***Enhanced Preferred***	***Normal Preferred***
59	100%	0%
23	92.9%	7.1%
10	85.7%	14.3%
3	78.6%	21.4%
2	71.4%	28.6%
2	64.3%	35.7%
1	57.1%	42.9%



Figure 11. Inverted YUV nighttime image looks like foggy daytime image.



Figure 12. Preprocessing algorithm utilizing CETM defogging.



Figure 13. Nighttime images before and after preprocessing.

4.2 Experiment on Stereo Pretrained Model with Dataset

To demonstrate our analysis, we utilize the Oxford RobotCar Dataset [12][13]. Our test set contains 6,001 randomly selected nighttime street view images from the 2014-11-14 dataset with images covering a variety of scenarios including glare, motion blur, noise, and overexposed lights. The corresponding ground truth disparities are obtained as detailed in Section 3.1. All test images are preprocessed or enhanced with CETM defogging algorithm described in Section 2.2 with the

procedure in Section 4.1 to produce clearer, brighter images. Then, we test our non-enhanced and enhanced images on depth estimation algorithms to evaluate if our method outperforms the original method which has no preprocessing.

For this particular test, we compare the error of two types of depth estimation. In the first input, the car hood is cropped from normal test images, enhanced test images, and ground truth disparity. In the second input, the car hood is not cropped from the images or ground truth. In Section 4.2.3, we test whether keeping the car hood, which can give noisy results, or cropping it out gives higher accuracy.

4.2.1 Stereo Depth Estimation with Dataset

For stereo depth estimation, we choose Stereo-Consistent Cyclic Translations [10], an algorithm created specifically for nighttime images. It is trained using a generative adversarial network (GAN). We run stereo images from the left and right camera through the pretrained model and achieve the disparity results shown in Figure 14.

In the first image, there are two bicycles, a car, a traffic light. The shapes of the bicycles (lime green boxes) and traffic light (purple box) are sharper in the normal image depth though the enhanced depth covers a similar area . However, the car (blue box) is clearer in the enhanced depth which correctly estimates the back half of the car that is missing in the normal depth. Enhanced depth captures the tree (green box) branches and leaves in the top left that normal depth completely ignores. Both detect the street and pedestrians (red box) on the right well.

In the second image, the enhanced depth fills in the left tree trunk (lime green box) with proper depth while it is spotty in the normal depth. The traffic poles (purple boxes) on the left appear as blotches in the normal depth due to low visibility. However, individual poles can be seen

in the enhanced depth giving a cleaner, more accurate result. Both estimate the street and car (blue box) well. Both enhanced and normal detect the two buildings (red boxes) further down the left and right of the street, but enhanced depth detects the shape and size better.

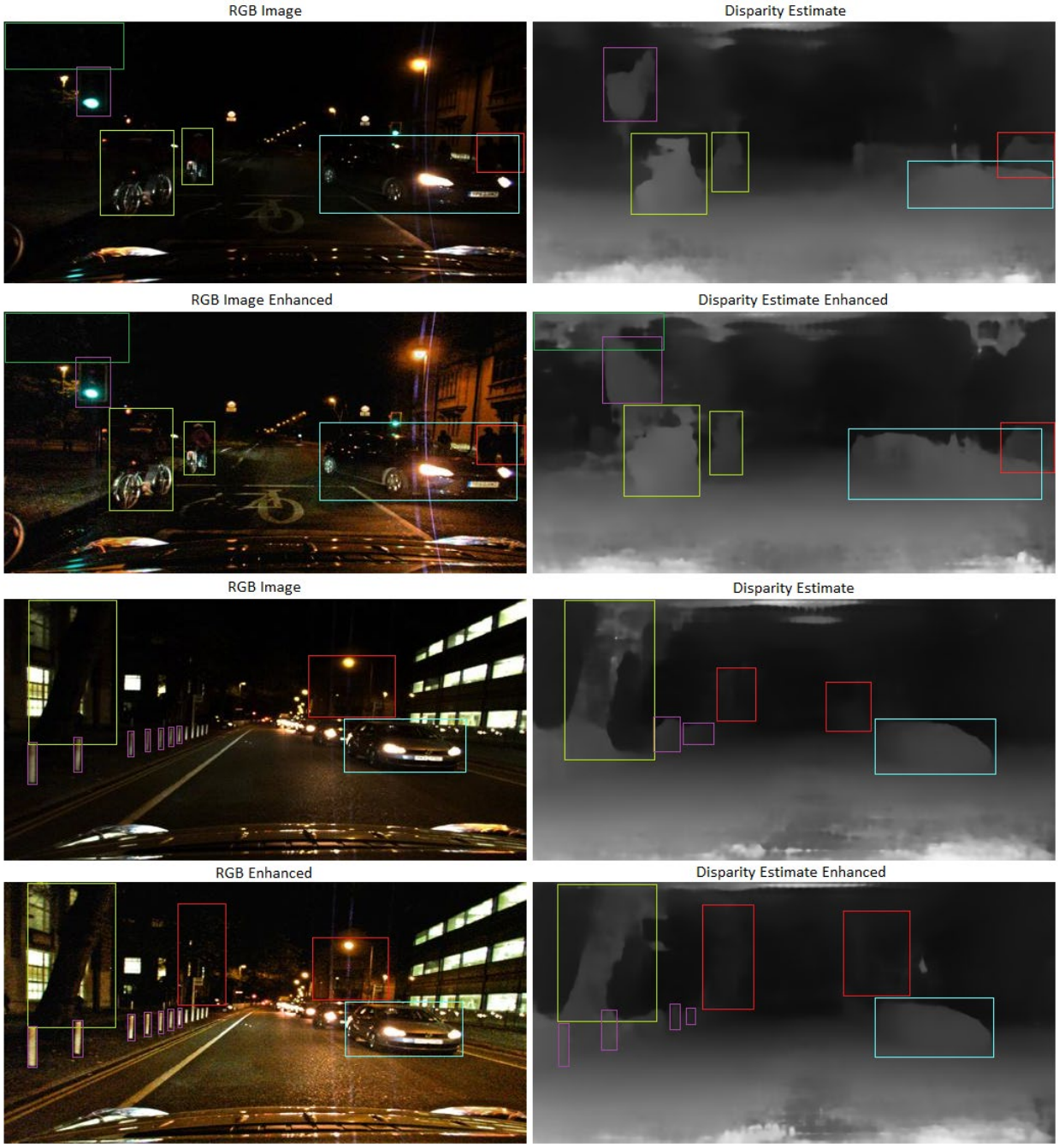


Figure 14. Estimated disparity for normal nighttime images and preprocessed nighttime images for stereo camera using Stereo-Consistent Cyclic Translations [10].

4.2.2 Error Evaluation Metric

We choose the disparity error to compare the performance of the proposed algorithm with conventional ones. We calculate the disparity error as the percentage of pixels that are k away from the true disparity where k is varied from 0 to 20 and k=0 indicates a correct disparity estimate while k=20 indicates large error. Since our ground truth is sparse, we only test non-zero ground truth pixels. The lower 10% of the ground truth and estimated disparity maps are cropped away as they are often covered by a car hood and return noisy results. Furthermore, the estimated disparity estimates' resolutions are resized to match the ground truth size of 512×256 using nearest neighbor. These steps are included in Figures 17 and 18 where the car hood is removed, and the disparity maps are resized.

A good disparity estimation algorithm should yield lower percentage error and higher percentage accuracy. We use an accumulated pixel percentage accuracy (APPA) metric for this experiment. Let *P(k)* be the percentage pixel accuracy for a given k, then the accumulation function is defined as

$$Y(k) = Y(k-1) + P(k)$$

which yields $Y(0) = P(0), Y(1) = P(0) + P(1), Y(2) = P(0) + P(1) + P(2)$, and so on.

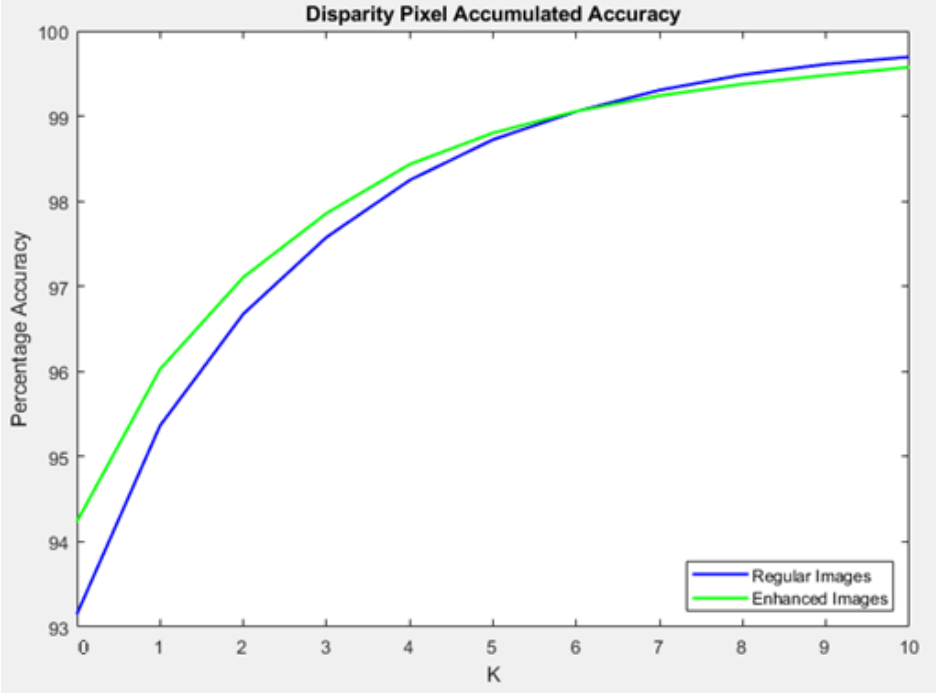
4.2.3 Stereo Results for Dataset

APPA results for the stereo depth estimation experiments are shown in Figure 15 and 16 for images including the car hood and with the car hood cropped out. The error and correctness results are averaged for all 6,001 images. Closer examination of individual image results is provided in later sections.

When the car hood is included in images and ground truth disparity, the accuracy of the normal and enhanced nighttime data begins at 93.2% and 94.2%. Enhanced disparity accuracy begins around 1.1% greater than non-enhanced disparity accuracy. The graph increases until both converge at k=6 after which normal images have greater accumulated accuracy than enhanced images from k=6-10. In total, 89.14% of pixels in normal images were estimated correctly while 90.31% of pixels in enhanced images were estimated correctly.

With the car hood cropped out of the images and ground truth disparity, the accumulated accuracy for normal and enhanced images begins at 93.4% and 94.3%. They begin approximately 1.2% apart and gradually increase until they converge at k=10. In this experiment, the accumulated accuracy of normal images is never greater than that of enhanced images.

In both cases with or without the car hood, images enhanced with our preprocessing have greater overall accuracy for depth estimation using Stereo-Consistent Cyclic Translations. The only exception is after k=6 for images with the car hood included. Cropping out the car hood is the better than leaving it in as it results in greater starting accuracy for k=0-10, similar ending accuracy, and always performs better with our preprocessed, enhanced images than the normal images. In addition the preprocessed images yield greater correct estimates than non-enhanced images. We conclude that nighttime images enhanced with the proposed preprocessing achieve greater disparity accuracy than the original state-of-the-art algorithm.

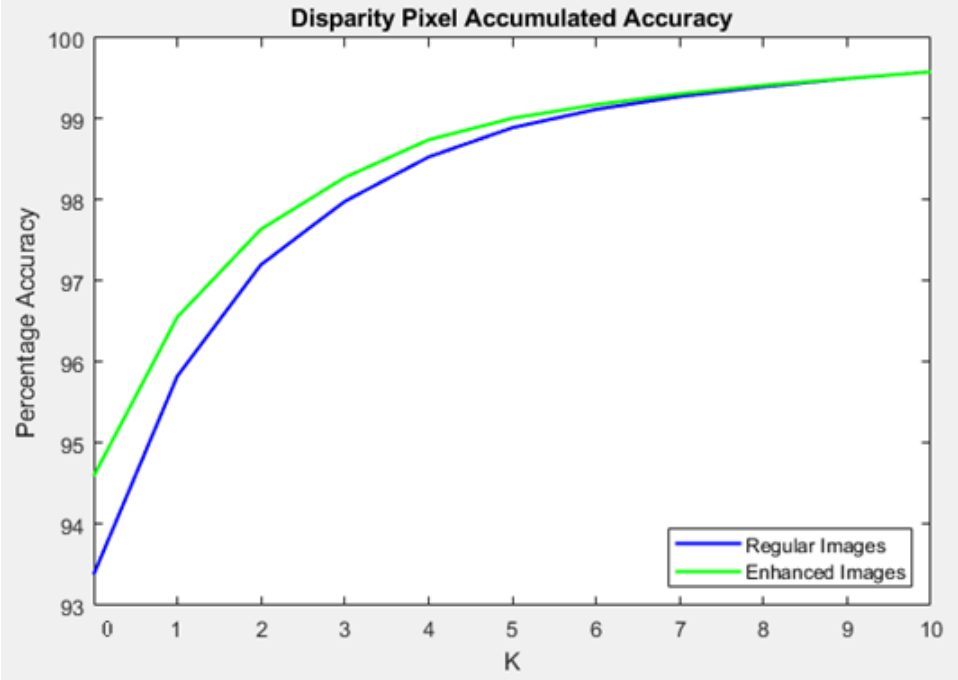


(a)

correctK_enhanced_avg	0.9031
correctK_normal_avg	0.8914

(b)

Figure 15. (a) Accumulated pixel percentage accuracy for stereo depth estimation with pretrained model on 6,001 nighttime images in the Oxford dataset without cropping the front car hood. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.



(a)

correctK_enhanced_avg	0.9162
correctK_normal_avg	0.9032

(b)

Figure 16. (a) Accumulated pixel percentage accuracy for stereo depth estimation with pretrained model on 6,001 nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.

4.3 Experiments on Monocular and Stereo Pretrained Models

Next, results of individual images from monocular and stereo depth estimation algorithms are examined. Since the proposed preprocessing method improved accuracy on the entire dataset for stereo depth estimation, it may work well with monocular cameras too. If it does, the preprocessing method is more robust to monocular and stereo depth estimation applications.

4.3.1 Monocular Depth Estimation with Single Images

For monocular depth estimation from a single image, we choose monodepth2 [3], a prominent algorithm which uses fully convolutional networks. We run single images from the left stereo camera through the mono+stereo_1024×320 model pretrained on daytime images. The output depth is converted to disparity so it can be compared to ground truth disparity later. Some sample results are shown in Figure 17.

In the first image, there are three cars, a building, and nothing is visible on the left side of the road though there could be a tree there. The normal depth detects a close object (blue box) where a tree could be located, but it has incorrect depth values as the tree shouldn't be so close and the gradient of decreasing depth does not resemble a tree. Enhanced depth does not detect a tree. However, it correctly detects the three cars (yellow boxes) while normal depth only detects two where the smaller car is barely detected. Both detect the street and building (purple box) well. The curb (green box) is detected incorrectly in normal depth to be much closer than it actually is, but enhanced depth is quite accurate.

In the second image, the difference is more subtle because the original image is somewhat bright already. The motorcycle (blue box) is detected much larger than it really is in normal depth while enhanced depth sizes it correctly and has predicts a more accurate shape. The traffic sign (purple box) is also detected more accurately in enhanced depth as a rectangle while it looks like a hexagon in normal depth. Both have difficulty detecting the car (yellow box) on the right, but detect the side door well. Only normal depth detects a sliver of the lamp post (green box) on the left.



Figure 17. Estimated disparity for normal nighttime images, preprocessed nighttime images, and sparse ground truth for monocular camera using monodepth2 [3].

4.3.2 Stereo Depth Estimation with Single Images

For stereo depth estimation, we choose Stereo-Consistent Cyclic Translations [10], an algorithm created specifically for nighttime images. It is trained using a generative adversarial network (GAN). We run stereo images from the left and right camera through the pretrained model and achieve the disparity results shown in Figure 18.

In the first image, there are three cars, a building, and potentially a tree on the left. The tree (blue box) is barely visible causing normal depth estimation to be noisy and erratic as expected. It has sudden drops in depth value and a lot of noise (bottom left of blue box). Once the image is

preprocessed, the depth of the tree is much smoother which is more accurate to a tree. Unlike monocular depth, in stereo depth, both normal and enhanced depth detect all three cars (yellow boxes). Additionally, they both detect the building (purple box) is accurately.

In the second image, there are two cars, a motorcycle, traffic light, and lamp post. The thin lamp post (green box) is barely detected in normal depth with only two small slivers, but it is much more visible in enhanced depth with two large slivers detected. Normal and enhanced depth accurately detect the car (yellow box) on the right and the motorcycle (blue box). Although, both detect the center car (yellow), normal depth has a jump in gradient from the car's trunk to back windshield which is inaccurate. Enhanced depth of the center car is smoother.

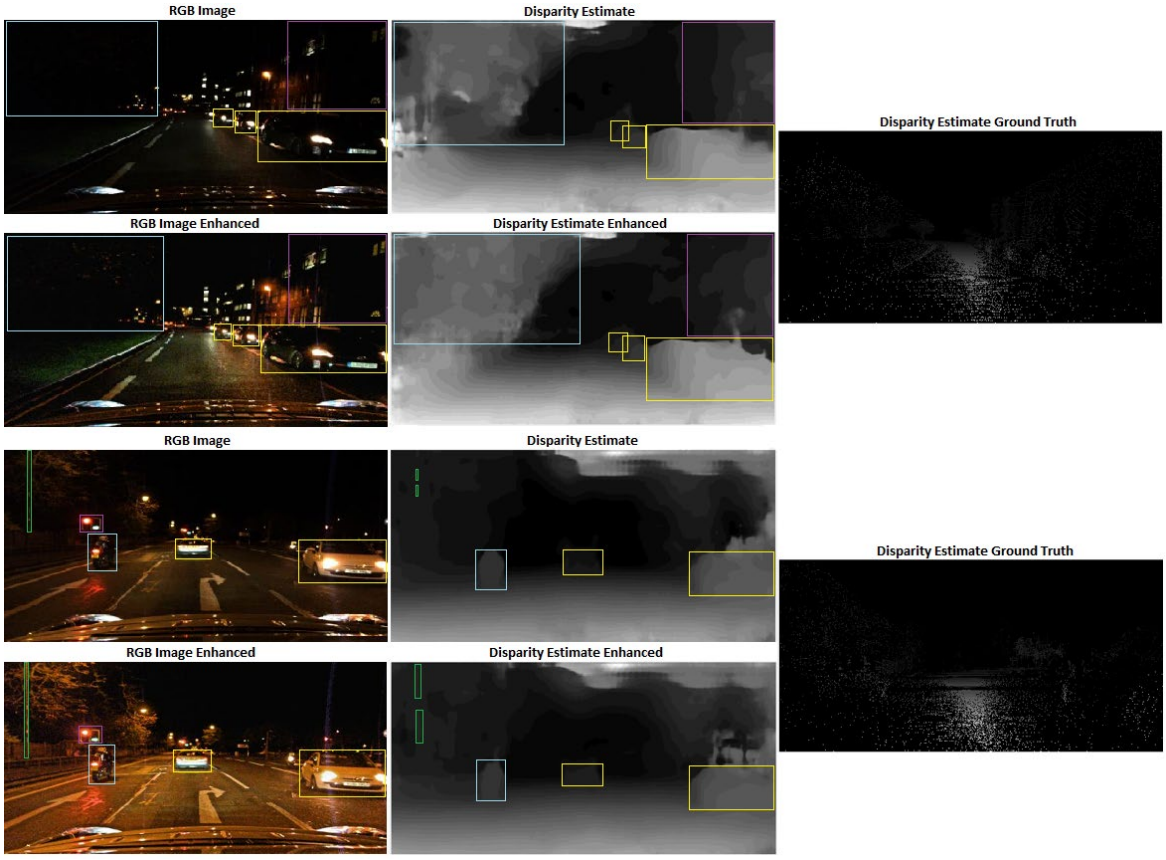


Figure 18. Estimated disparity for normal nighttime images, preprocessed nighttime images, and sparse ground truth for stereo cameras using Stereo-Consistent Cyclic Translations [10].

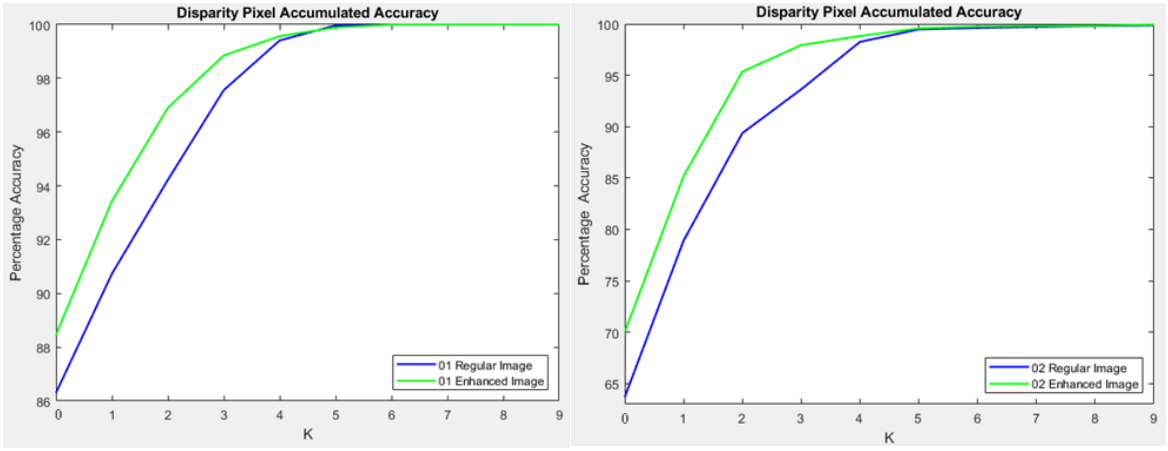
4.3.3 Monocular and Stereo Results for Single Images

APPA results for these experiments are shown below. For the monocular case in Figure 19, the accuracy of the first normal and enhanced nighttime sample begins around 86.5% and 89% and enhanced disparity accuracy remains 2-3% greater than non-enhanced disparity accuracy until they converge at k=4 which is close to 100% accuracy. This occurs because there are almost no incorrectly classified pixels at higher k. Here, depth estimation correctly predicts 86.30% of normal image pixels and 88.47% on enhanced image pixels. In the second sample, normal and enhanced accuracy begin at 63% and 70%. They rise steadily with enhanced depth accuracy 5-8% greater than non-enhanced disparity accuracy until k=4. Here, depth estimation correctly predicts 63.59% of normal image pixels and 69.76% on enhanced image pixels.

For the stereo case in Figure 20, the accuracy of the first normal and enhanced nighttime samples begins around 91% and 93.5%. Enhanced disparity accuracy remains 1-2% greater than non-enhanced disparity accuracy. The graph increases until k=6 at which point the accumulated accuracy has almost reached 100% at this point. Here, depth estimation correctly predicts 93.97% of normal image pixels and 94.11% on enhanced image pixels. In the second sample, normal and enhanced nighttime disparity accuracy begin at 91 % and 93%. Enhanced depth accuracy remains 1-2% greater than non-enhanced depth accuracy until k=4 when they converge. Here, depth estimation correctly predicts 92.41% of normal image pixels and 92.87% on enhanced image pixels.

In both monocular and stereo cases, images enhanced with our preprocessing have greater accuracy for monodepth2 and Stereo-Consistent Cyclic Translations. Monocular depth in particular performs much better with our preprocessing than without while stereo depth has slightly greater accuracy. Depth estimation has greater accuracy on enhanced images in every test. We conclude

that nighttime images enhanced with the proposed preprocessing achieve greater disparity accuracy than the original state-of-the-art algorithms.



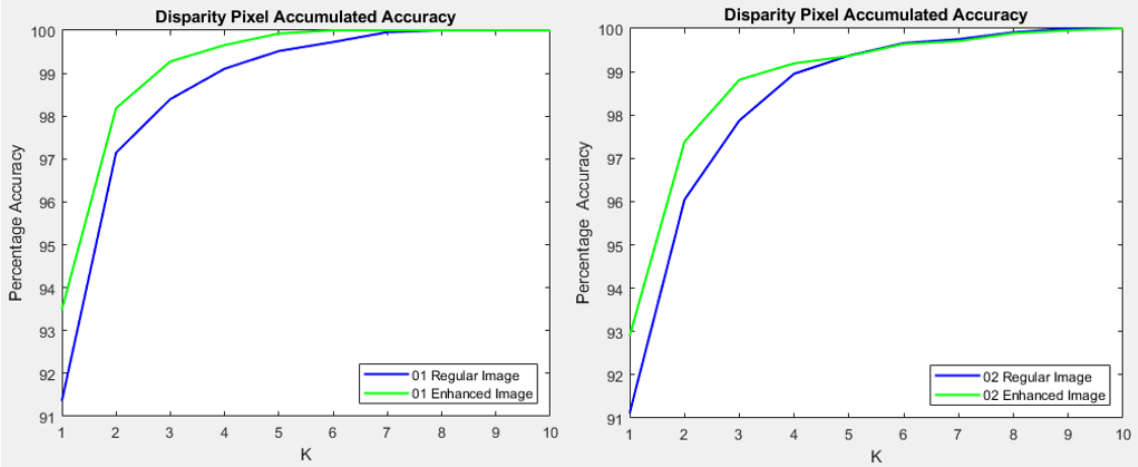
(a)

correctK0_est_groundTruth	0.8630
correctK0_estEnhanced_groundTruth	0.8847

correctK0_est_groundTruth	0.6359
correctK0_estEnhanced_groundTruth	0.6976

(b)

Figure 19. (a) Accumulated pixel percentage accuracy for monocular depth estimation with pretrained model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.



(a)

correctK0_est_groundTruth	0.9387	correctK0_est_groundTruth	0.9241
correctK0_estEnhanced_groundTruth	0.9411	correctK0_estEnhanced_groundTruth	0.9287

(b)

Figure 20. (a) Accumulated pixel percentage accuracy for stereo depth estimation with pretrained model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.

4.4 Experiment on Self-Trained Stereo Model

In the last section, we showed that our proposed preprocessing algorithm improves depth estimation for low-light and nighttime images. It yields a more accurate depth estimation for both monocular and stereo-based depth estimation algorithms. Next, we train a custom model using Depth Estimation in Nighttime using Stereo-Consistent Cyclic Translations [10] on enhanced data to see if it will yield greater accuracy on input enhanced nighttime images. This requires stereo training and validation data. We prepare our dataset with images from the Oxford RobotCar Dataset [12][13] dataset covering a variety of scenarios including glare, motion blur, noise, and overexposed lights. For the training dataset, we randomly select 9,000 stereo daytime images from the 2014-07-14 dataset. Additionally, we randomly select 5,925 nighttime images from 2014-11-

14 and 2,975 from 2014-12-10. For our validation dataset, we randomly select 1,000 nighttime images from 2014-11-14. All nighttime data is preprocessed with the CETM enhancement algorithm described in Section 4.1 to defog, increase contrast, and brighten the images. As our validation data requires ground truth disparity for training, we calculate this by projecting Lidar scans into the camera [5] as detailed in Section 3.1.

4.4.1 Training Parameters

To train our model, we crop all training and validation images to 256×512. We train for 26 epochs and set the following parameters for our stereo GAN. Initial learning rate for the Adam [20] optimizer starts with $lr = 0.0002$ for the first 20 epochs of training and linearly decays to 0 in the last 6 epochs. It has momentum weights $\beta_1 = 0.5$, $\beta_2 = 0.999$ and the cycle-consistency [21] loss multiplier is $\lambda_{cyc} = 10.0$. The disparity reconstruction loss multiplier is $\lambda_{ste} = 0.05$. The batch size is 4 and each translation is buffered with the 50 previous pairs to mitigate spikes in the model's loss function. Maximum disparity is set to 48 as further objects are not as important in autonomous driving.

4.4.2 Stereo Depth Estimation with Single Images

After preprocessing the night images, we train the model for 26 epochs on daytime and enhanced nighttime images which takes 5 days on one Nvidia GeoForce 1080Ti GPU. After our custom model is trained with the preprocessed nighttime images, we run stereo images from the left and right camera through the trained model with the process shown in Figure 21 where $\hat{g}_N^*$, $\hat{g}_D^*$, and $\hat{f}_N^*$ are the optimized weights we trained as detailed in Section 2.3.1. The disparity

estimation results are shown in Figure 22 where both normal and preprocessed images are passed through the model. Visually, the resulting disparity maps look fairly similar for both.

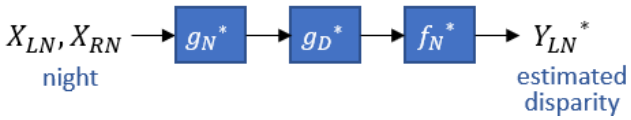


Figure 21. Test custom disparity estimation model with nighttime images.

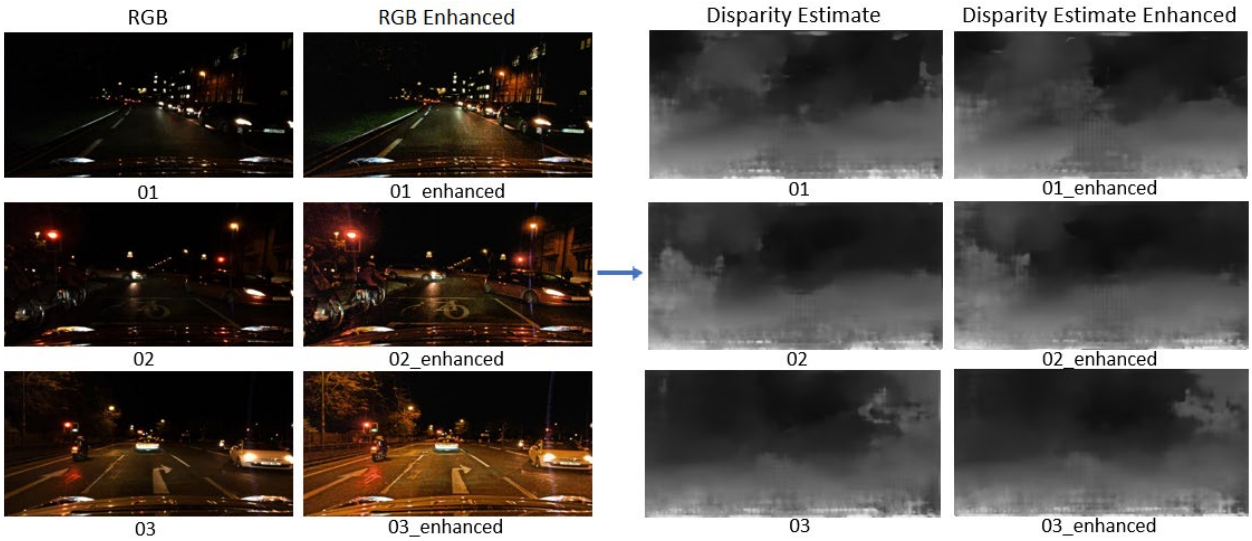


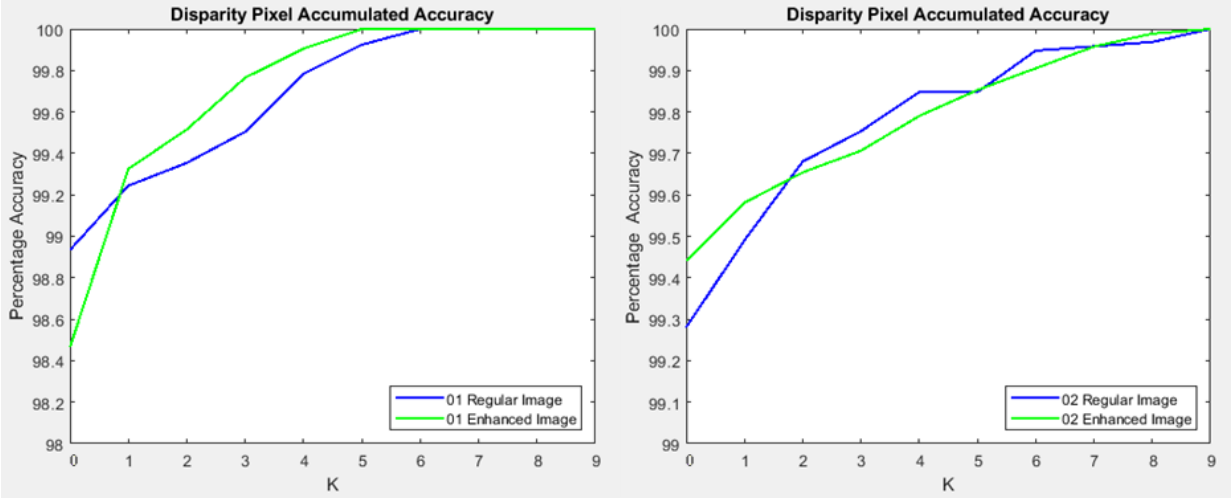
Figure 22. Estimated disparity on custom model with nighttime images, preprocessed nighttime images, and sparse ground truth using Stereo-Consistent Cyclic Translations [10].

4.4.3 Stereo Results for Single Images

As with the pretrained models, we evaluate our custom model disparity output with the APPA metric defined in Section 4.2.2. APPA results for these experiments are shown in Figure 23. In the first test sample, the accuracy of the normal and enhanced nighttime images begins around 98.95% and 98.5%. Enhanced depth accuracy remains ≤0.3% greater than non-enhanced depth accuracy as the graph increases from $k=1$-6 where both accumulated accuracy reach 100%. Here, depth estimation correctly predicts 98.92% of normal image pixels and 98.46% on enhanced image pixels. In the second sample, the accuracy of normal and enhanced nighttime images begins at

99.28% and 99.45%. Enhanced Depth is ≤0.15% greater than non-enhanced depth accuracy until k=2 when they converge. Here, depth estimation correctly predicts 99.28% of normal image pixels and 99.46% on enhanced image pixels.

With the model trained on images enhanced with our preprocessing, preprocessed input images yield depth with slightly greater accuracy than images without our preprocessing. However, the starting depth accuracy of our custom model is significantly higher than that of the pretrained model which is not trained on enhanced images by 5-6%. Additionally, depth estimation has higher accuracy on enhanced compared to normal images in every test except the self-trained model. We conclude that nighttime images enhanced with the proposed preprocessing achieve greater disparity accuracy than the original state-of-the-art algorithms and our self-trained model is more accurate than pretrained models.



(a)

correctK0_est_groundTruth	0.9892	correctK0_est_groundTruth	0.9928
correctK0_estEnhanced_groundTruth	0.9846	correctK0_estEnhanced_groundTruth	0.9946

(b)

Figure 23. (a) Accumulated pixel percentage accuracy for stereo depth estimation with custom model on single nighttime images in the Oxford dataset with the front car hood cropped out. (b) Percentage of normal and enhanced pixels estimated correctly or k=0 in APPA.

Chapter 5

Enhanced Object Detection for Low-Light

5.1 Experiment on Pretrained Object Detection Model

Given a single RGB image of a nighttime scene, our goal is to detect objects such as cars, trucks, bicycles, people, stop signs, traffic lights, and more. Each detection has a corresponding confidence percentage of its accuracy. Additionally, the bounding box of each detected object's location in the image is returned. The proposed approach consists of two steps. First, it uses the CETM defogging method to preprocess and enhance the image. Then, it uses a CNN model to detect objects.

To demonstrate our analysis, we utilize the Berkeley Deep Drive dataset. Our test set contains 7,130 nighttime images including some with motion blur, weird camera angles, dirty windshields, or well-lit environments. Each image has a corresponding image enhanced with the CETM defogging algorithm described in Section 2.2 with the procedure described in Section 4.1. This results in clearer, brighter images. In addition, there are ground truth bounding box labels for objects in each image. Then, we test our non-enhanced and enhanced images on the YOLOv5 RGB depth estimation algorithm to evaluate if our method outperforms the original method which has no preprocessing.

5.1.1 Object Detection with Dataset

For object detection, we choose YOLOv5, an algorithm created for real-time object detection. It is trained using a CNN. We test 7,130 single images on the pretrained YOLOv5s model and achieve disparity results as shown in Figure 24.



Figure 24. Objects detected in normal nighttime images and preprocessed nighttime images using YOLOv5 pretrained model.

In the first image, there are three cars (orange boxes) and a truck (green box) detected. The truck is not detected in the normal image, but detected with 0.28 confidence once the image is enhanced. However, the cars are detected with slightly higher confidence in the normal image. In the second image, there are eleven cars (orange boxes) and two traffic lights (turquoise boxes) detected. The enhanced image detects most cars with slightly higher confidence while the normal image has higher confidence detecting the traffic lights.

5.1.2 Error Evaluation Metric

To obtain an accurate measure of the object detection results, we utilize the mean average precision (mAP) accuracy metric. For each bounding box, it measures the overlap between the predicted and ground truth bounding box which is called IoU (intersection over union). A set IoU threshold value determines the precision which depends on the number of true or false positive detections.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

If the IoU of a bounding box is higher than the IoU threshold, it is considered a true positive. Otherwise, it is a false positive. The x-axis of the precision-recall graph is recall which measures how well a model finds all the ground truths objects and is determined as

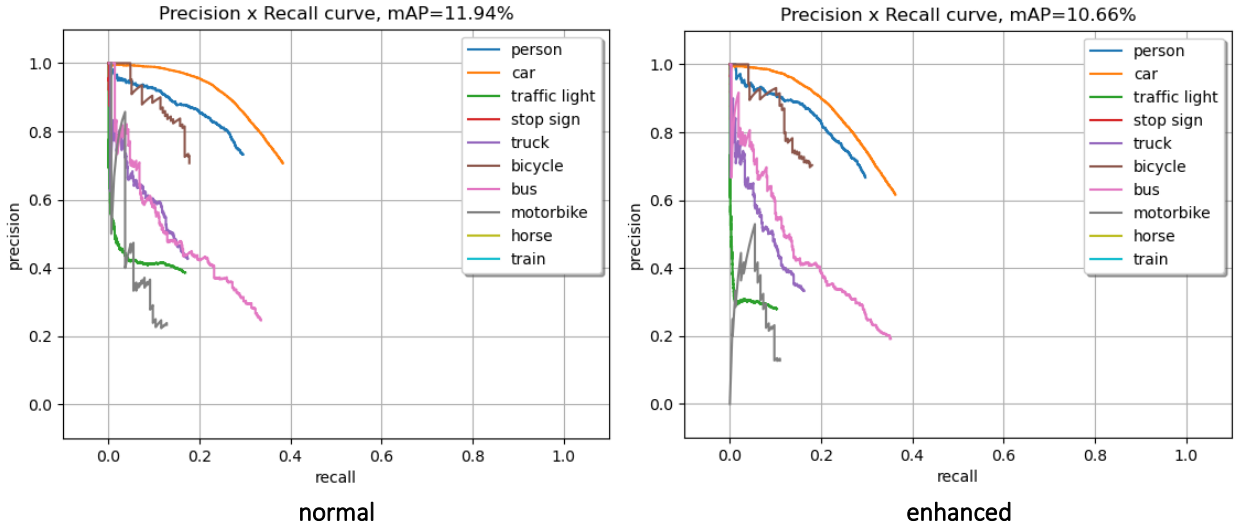
$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Once the precision-recall curve is obtained, the average precision (AP) is found by calculating the area under the curve. Mean average precision or mAP score is calculated by taking the mean AP over all classes. Higher mAP scores correlate to better object detection.

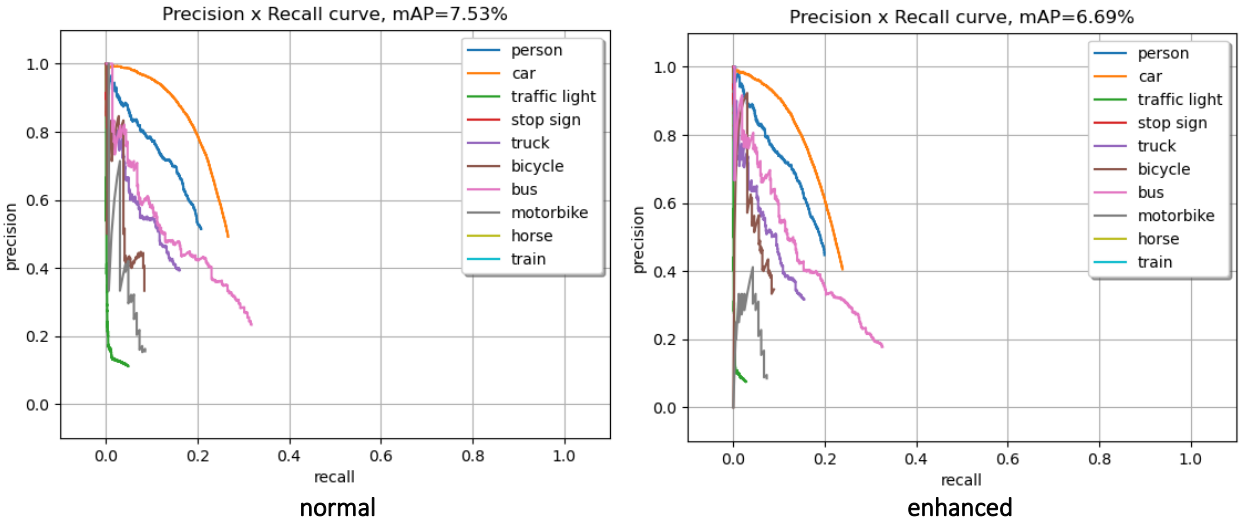
5.1.3 Object Detection Results for Dataset

We evaluate the pretrained model object detection results with the mAP metric defined in Section 5.1.2. The results for these experiments are shown in Figure 25 for IoU=0.5, 0.7, 0.85, and 0.93 where the higher thresholds are stricter on what is detections are considered true positives. Normal images have a higher mAP score than enhanced images for every IoU tested. The percentage difference between normal and enhanced mAP scores decreases from 1.28% to 0.22%

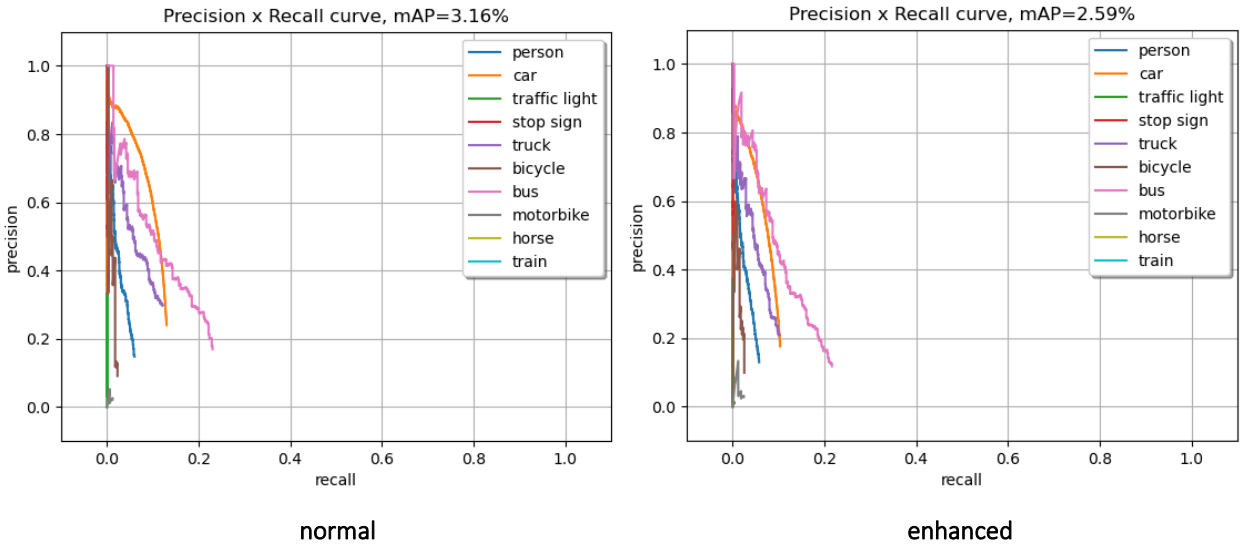
as IoU increases. Cars are detected most accurately in all test cases and then people, buses, and bicycles in no particular order. We conclude that nighttime images enhanced with the proposed preprocessing do not achieve greater object detection accuracy than normal images with the pretrained YOLOv5 model.



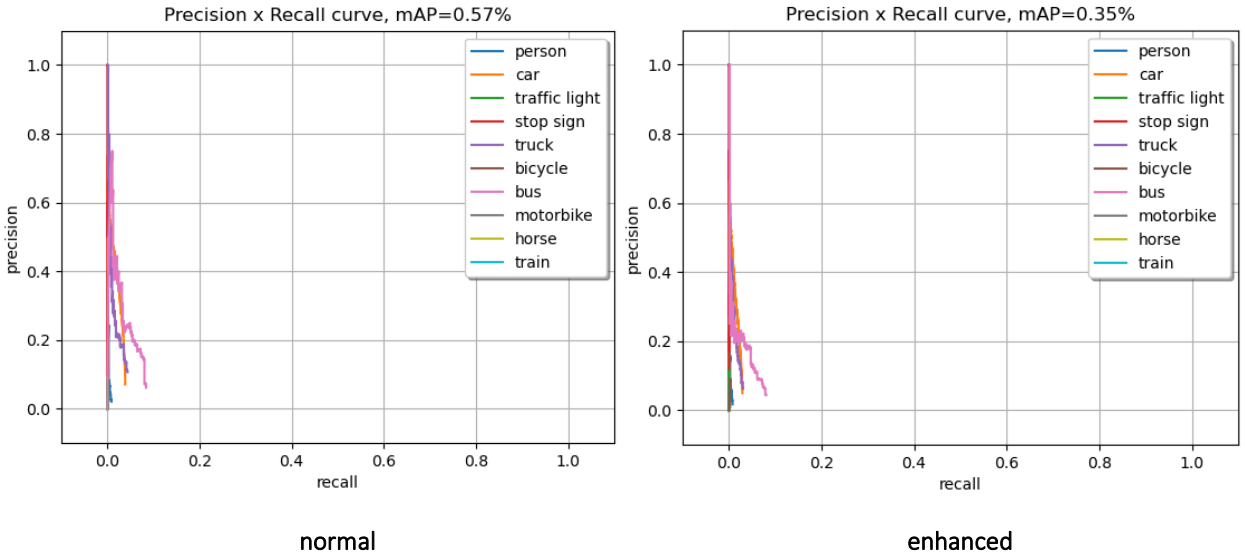
(a)



(b)



(c)



(d)

Figure 25. Mean average precision (mAP) evaluation on YOLOv5 nighttime object detection for pretrained model with (a) IOU=0.5 (b) IOU=0.7 (c) IOU=0.85 (d) IOU=0.93.

5.2 Experiment on Self-Trained Object Detection Model

In the final section, we showed that our proposed preprocessing algorithm does not improve object detection for low-light and nighttime images. It yields lower accuracy on the YOLOv5 algorithm evaluated with mAP. This is likely because the pretrained model is trained

primarily on daytime images. Next, we train a custom YOLOv5 model by training additional enhanced data to the pretrained model to see if it yields higher accuracy on input enhanced nighttime images. This requires monocular training and validation data. We prepare our dataset with 1,200 nighttime images from the Berkeley Deep Drive dataset. All nighttime data is preprocessed with the CETM enhancement algorithm described in Section 4.1 to defog, increase contrast, and brighten the images. We use our training data as validation data.

5.2.1 Object Detection with Dataset

After preprocessing the night images, we train for an additional 300 epochs on the base YOLOv5s model on one Nvidia GeoForce 1080Ti GPU. After our custom model is trained, we select the epoch with the best model. Then, we run normal and enhanced Berkeley Deep Drive test images through the trained model. The disparity estimation results are shown in Figure 26 where both normal and preprocessed images are passed through the model. Visually, the self-trained model performs much better than the pretrained model. It is able to detect stop signs and detects cars, traffic lights, and trucks with 0-19% higher confidence than the pretrained model.

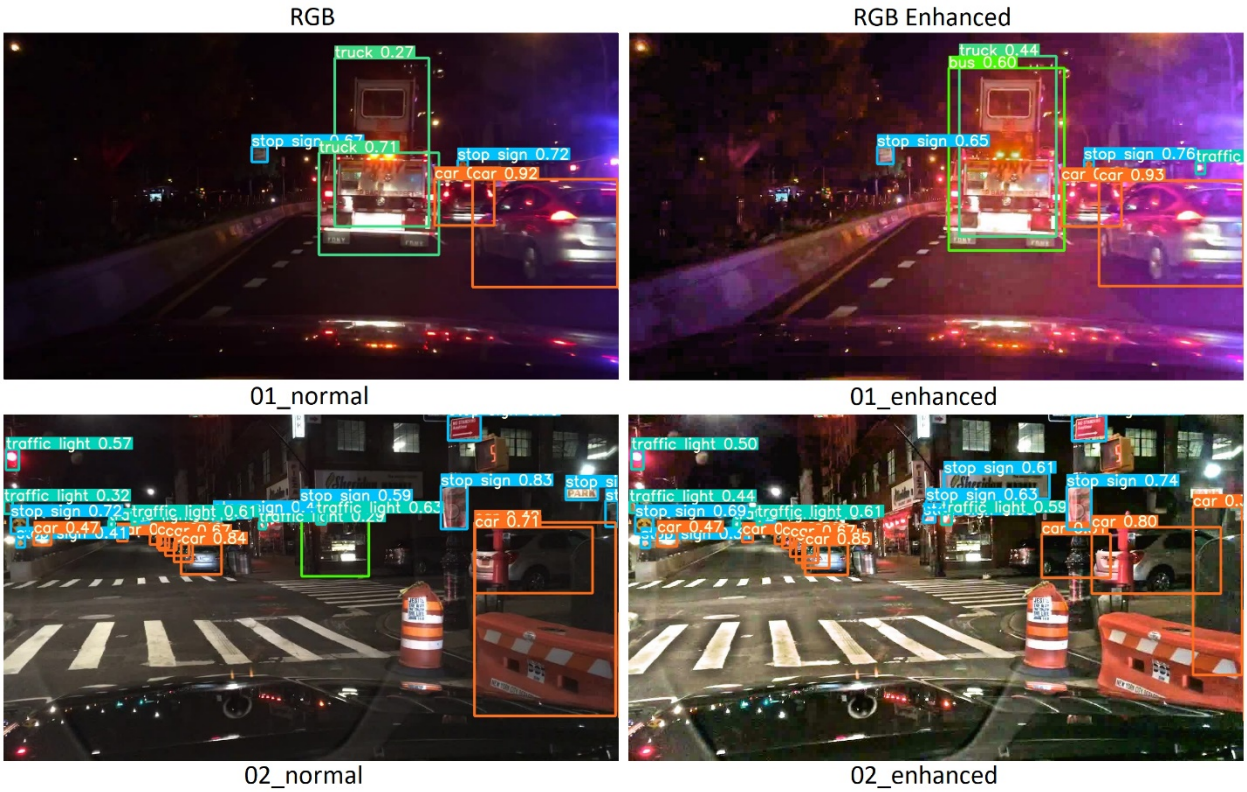


Figure 26. Objects detected in normal nighttime images and preprocessed nighttime images using YOLOv5 self-trained model.

In the first image, there are two cars (orange boxes), a truck (green box), and two stop signs (blue) detected. Cars, trucks, and stop signs are detected with 1%, 4-8%, and 27% higher accuracy for the enhanced image than the normal image. The firetruck is not featured much in training data causing it to be detected as two separate trucks or a truck and bus. In the second image, there are 10 cars (orange boxes) detected, seven stop signs (blue boxes), and five traffic lights (turquoise boxes). Cars are detected with 1-9% higher for the enhanced images than the normal image. Traffic lights and stop signs are equally detected with higher confidence in the normal and enhanced image.

5.2.2 Object Detection Results for Dataset

We evaluate the pretrained model object detection results with the mAP metric defined in Section 5.1.2. The results for these experiments are shown in Figure 27 for IoU=0.5, 0.7 where the higher threshold is stricter on what is detections are considered true positives. Enhanced images have higher mAP scores than those of normal images for every IoU tested. The percentage difference between normal and enhanced mAP scores decreases from 6.81% to 5.04% as IoU increases. Cars are detected most accurately in all test cases and then stop signs, trucks, and buses in no particular order. We conclude that nighttime images enhanced with the proposed preprocessing achieve higher object detection accuracy than normal images with the self-trained YOLOv5 model.

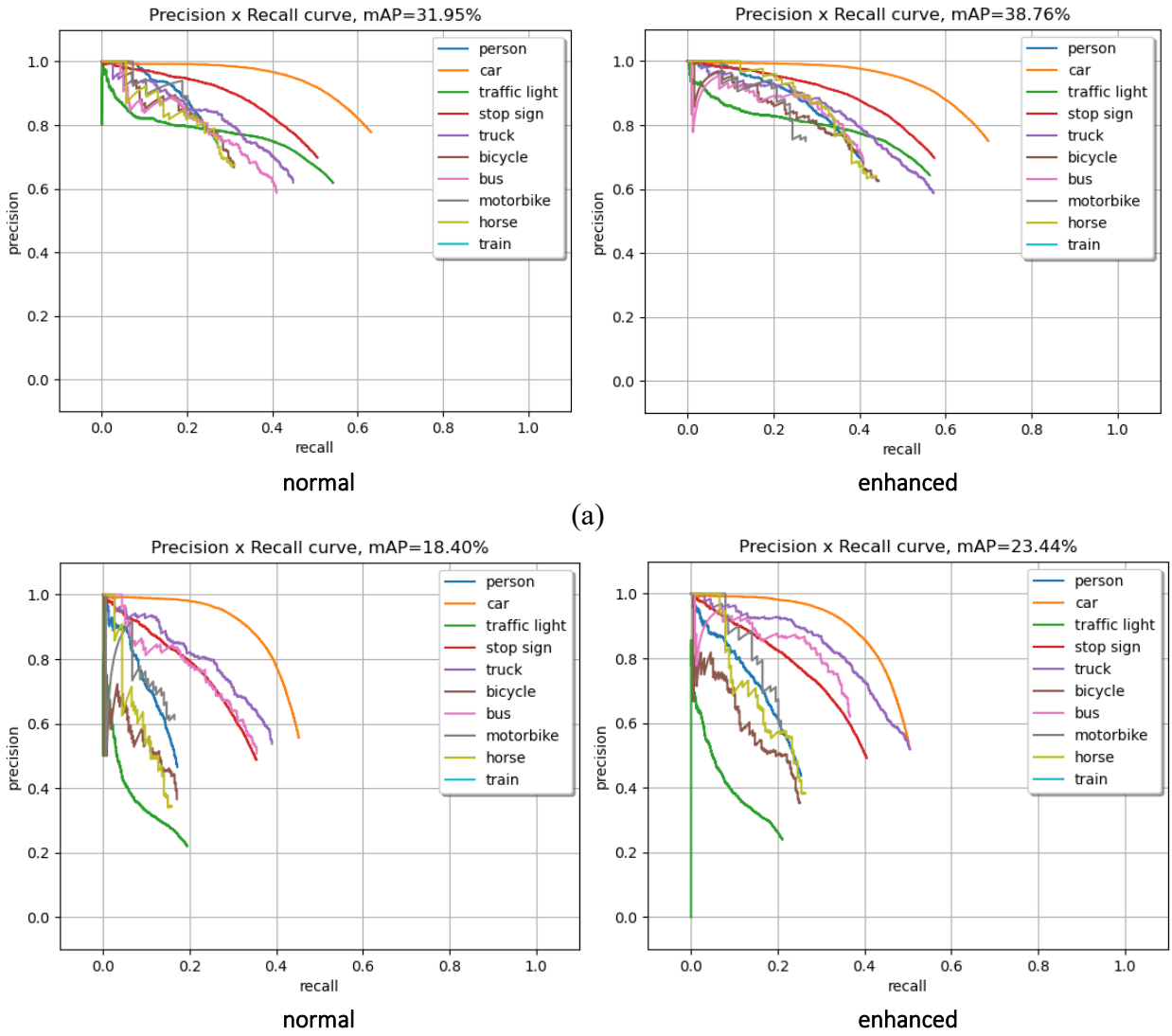


Figure 27. Mean average precision (mAP) evaluation on YOLOv5 nighttime object detection for self-trained model with (a) IOU=0.5 (b) IOU=0.7.

Chapter 6

Conclusion

We propose a preprocessing algorithm to improve depth estimation and object detection for low-light and nighttime images. The preprocessing algorithm consists of inverting the low-light image and applies a defogging and turbulence mitigation algorithm. Additionally, we train a custom stereo depth estimation model for enhanced images. Both pretrained and custom models are tested with low-light street view images from a variety of scenarios including glare, dark objects, noise, glow, and overexposed light. Resulting depth is evaluated with the accumulated pixel percentage accuracy metric. We also train a custom YOLOv5 object detection model with low-light street view images. The resulting detections are evaluated with the mean average precision metric. The proposed method yields more accurate depth estimation for both monocular and stereo-based depth estimation algorithms. It also improves object detection performance for nighttime images.

This design improves the performance of current nighttime disparity estimation, but low-light depth estimation is still a largely untouched field. Further research towards low-light applications includes object detection, semantic segmentation, or 3D reconstruction to improve safety for autonomous vehicles particularly when encountering pedestrians, vehicles, and guardrails.

Bibliography

[1] Yong Seok Heo, Kyoung. Mu Lee, Sang Uk Lee, "Joint Depth Map and Color Consistency Estimation for Stereo Images with Different Illuminations and Cameras," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 5, pp 1094-1106, May 2013, doi: 10.1109/TPAMI.2012.167.

[2] Aashish Sharma, Lionel Heng, Loong-Fah Cheong, Robby T. Tan, "Nighttime Stereo Depth Estimation using Joint Translation-Stereo Learning: Light Effects and Uninformative Regions," in International Conference on 3D Vision (3DV), pp. 23-31, November 24-28, 2020, doi: 10.1109/3DV50981.2020.00012.

[3] Clement Godard, Oisin Mac Aodha, Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Conference on Computer Vision and Pattern Recognition (CVPR), pp. 270-279, 2017.

[4] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," in European Conference on Computer Vision (ECCV), pp 740-756, 2016.

[5] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, "Pseudo-lidar from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Diving," in Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8445-8453, 2019.

[6] Cherian Varghese, Umesh Shankar, "Passenger Vehicle Occupant Fatalities by Day and Night – A Contrast," in National Highway Traffic Safety Administration (NHTSA), issue HS-810 637, pp. 1-8, May 2007.

[7] *Night Blindness (Nyctalopia)*, in Cleveland Clinic, December 2020. Accessed on: June 2021. [Online]. Available: https://my.clevelandclinic.org/health/symptoms/10118-night-blindness-nyctalopia

[8] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv: 2004. 10934, 2020.

[9] Kristofor B. Gibson, Truong Q. Nguyen, "An Analysis and Method for Contrast Enhancement Turbulence Mitigation Methods," in IEEE Transactions on Image Processing, vol. 23, no. 7, pp. 3179-3190, July 2014, doi: 10.1109/TIP.2014.2328180.

[10] Aashish Sharma, Robby T. Tan, Loong-Fah Cheong, "Depth Estimation in Nighttime using Stereo-Consistent Cyclic Translations," arXiv e-prints: arXiv-1909, 2019.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Nets," in Advances in Neural Information Processing Systems, vol. 27, pp. 2672-2680, 2014.

[12] Will Maddern, Geofrrey Pascoe, Chris Linegar, Paul Newman, "1 year, 1000km: The Oxford Robotcar Dataset," in The International Journal of Robotics Research, vol. 36, no. 1, pp 3-15, 2016.

[13] Dan Barnes, Matthew Gadd, P. Murcutt, Paul Newman, Ingmar Posner, "The Oxford Radar Robotcar Dataset: A Radar Extension to the Oxford Robotcar Dataset," in IEEE International Conference on Robotics and Automation (ICRA), pp. 6433-6438, 2020, doi: 10.1109/ICRA40945.2020.9196884.

[14] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, Luc Van Gool, "Night-to-Day Image Translation for Retrieval-based Localization," in International Conference on Robotics and Automation (ICRA), pp. 5958-5964, 2019, doi: 10.1109/ICRA.2019.8794387.

[15] Jia-Ren Chang, Yong Sheng Chen, "Pyramid Stereo Matching Network," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5410-5418, 2018.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, "Image-To-Image Translation with Conditional Adversarial Networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp1125-1134, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine learning Intelligence, vol. 37, no. 9, pp. 1904-1916, Sept. 1, 2015, doi: 10.1109/TPAMI.2015.2389824.

[19] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, Adam Bry, "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in IEEE International Conference on Computer Vision (ICV), pp. 66-75, 2017.

[20] Jimmy Ba, Diederik P. Kingma, "Adam: A Method for Stochastic Optimization," in International Conference on Learning Representations (ICLR), vol. 5, pp. 346–361, 2014.

[21] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in International Conference on Computer Vision (ICCV), pp. 2223-2232, 2017.

[22] Moritz Menze and Andreas Geiger, "Object Scene Flow for Autonomous Vehicles," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3061-3070, June 2015, 10.1109/CVPR.2015.7298925.

[23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, Rob Fergus. "Indoor Segmentation and Support Inference from RGBD Images," in Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, "Computer Vision" – ECCV 2012, pp. 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[24] David Eigen, Christian Puhrsch, Rob Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," in "Advances in Neural Information Processing Systems," 27, pp. 2366–2374. Curran Associates, Inc., 2014.

[25] Tinghui Zhou, Matthew Brown, Noah Snavely, David G Lowe, "Unsupervised Learning of Depth and Ego-Motion from Video," in Proceedings of the IEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1851-1858, 2017.

[26] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, Simon Lucey, "Learning Depth From Monocular Videos Using Direct Methods," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2022-2030, 2018.

[27] Nikolai Smolyanskiy, Alexey Kamenev, Stan Birchfield, "On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1007-1015, 2018.

[28] Sameh Khamis, Sean Fanello, Christoph, Rhemann, Adarsh Kowdle, Julien Valentin, Sharam Izadi, "StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth

Prediction," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 573-590, 2018.

[29] Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, Sofien Bouaziz, "HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14362-14372, 2021.

[30] Zhengyu Huang, Theodore B. Norris, Panque Wang, "ES-Net: An Efficient Stereo Matching Network," in arXiv preprint, arXiv: 2103.03922, 2021.

[31] Steven Seitz. "An Overview of Passive Vision Techniques." SIGGRAPH 2000 Course on 3D Photography, Course Notes, New Orleans, Louisiana, LA, USA, 1999.

[32] *What is Structured Light Imaging*, in Robotics Tomorrow, April 24, 2018. Accessed on: June 2018. [Online]. Available: https://www.roboticstomorrow.com/article/2018/04/what-is-structured-light-imaging/11821